Consumer's Guide to Research on STEM Education

March 2012

Iris R. Weiss

Horizon
RESEARCH, INC.

Consumer's Guide to Research on STEM Education was prepared with support from the National Science Foundation under grant number EHR-0445398. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Consumer's Guide to Research on STEM Education

Periodically the field summarizes what is known in particular areas, in some cases supplementing the findings from empirical research with the insights of experts in the area. Typically involving people with a variety of backgrounds and perspectives in the development process, these summaries not only document the state of knowledge at a given juncture, but also provide guidance to practitioners and policymakers based on what is currently known.[1]

But comprehensive efforts of this sort are time consuming and expensive, and there are many areas of interest to STEM educators where the available knowledge has not been compiled in a practitioner-friendly form. Rather, a mathematics/science supervisor, teacher, or other educator may find out about studies in presentations at professional association meetings, newsletters, or journals and want to learn more. This guide is intended to help consumers of research assess the quality and implications of both individual studies of interventions and research syntheses.[2]

The following sections address two key questions that should guide practitioners in reviewing research:

1. How much should I trust the findings?
2. What are the implications, if any, for my context?

Section 1: How much should I trust the findings?

A primary consideration in evaluating the trustworthiness of an article is whether *how* a particular study was conducted biased its results. Decisions about who was included in a study, what information was collected from them, and how that information was analyzed are all possible sources of bias. Consumers of research need to be on the lookout for potential bias, as claims of effectiveness are suspect if there are other credible explanations for the results.

Look to see if there is a comparison group, and for evidence that the treatment and comparison groups were similar prior to the study.

Studies that claim that interventions were responsible for changes are most believable when they included people who experienced the intervention (a "treatment group") and people who did not (a

[1] See for example: National Research Council. (2000). *How people learn*. Washington, DC: National Academy Press; National Council of Teachers of Mathematics (NCTM). (2010). *Principles and standards for school mathematics.* Reston, VA: Author; and Siegler, R., Carpenter, T., Fennell, F., Geary, D., Lewis, J., Okamoto, Y., Thompson, L., & Wray, J. (2010). *Developing effective fractions instruction for kindergarten through 8th grade: A practice guide* (NCEE #2010-4039). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

[2] This consumer's guide highlights some of the key issues involved in research on STEM education interventions. A companion guide, "Learning Together: A User-Friendly Tool to Support Research on STEM Education Interventions," intended to help people design research, provides additional information. (URL: http://www.mspkmd.net/papers/research_tool.pdf)

"comparison group").[3] In the absence of a comparison group, data showing that the participants scored higher after treatment than before are not very convincing, as the difference could well be due to something other than the treatment. In particular, one would expect students to score higher at the end of the year than they did at the beginning of the year, with or without the particular intervention that was investigated!

But not any comparison group will do; how groups are selected matters. For example, a study might have compared teachers who volunteered to participate in the treatment to teachers who did not, or included teacher leaders in the treatment group and "regular" teachers in the comparison group. Results from studies where there was *selection bias* should be viewed with a healthy dose of skepticism. Teacher background and experience, school contexts, demographics, and performance on any pre-intervention assessments are some important ways to compare treatment and comparison groups to determine if they are similar.

Look for evidence that the data collection procedures were sound.

Research articles typically include descriptions of data collection procedures, ideally not only what instruments were used, but also providing information that allows readers to judge whether the instruments were appropriate for the study participants and the research questions. For example, it would likely be inappropriate for a study of an intervention aimed at elementary teachers to use a content assessment developed for high school science teachers. Similarly, if the assessment used was created by the developer of the intervention, the study should provide assurances that the instrument did not cause an unfair advantage for the treatment group, for example by asking questions in the context of scenarios addressed in professional development activities.

The study should also describe how the researchers ensured that using those instruments would provide trustworthy information. Administration procedures that differed across participants could lead to biased results. For qualitative data in particular, look for information about the qualifications and/or training of researchers who collected the data when considering the likely quality of the data.[4]

Studies are also expected to provide information about their response rates, that is, how many of the people asked to provide information actually did so. If the response rate was low, especially if it was less than 50 percent, then there is a question about whether the people who chose to respond were representative of their group overall.[5]

It is important to note that differential patterns of response can bias the results either positively or negatively. For example, if a substantial number of the initially least prepared teachers in the

---

[3] Although not ideal from a research design perspective, comparisons between participants and an external standard can provide evidence of an intervention's impact. For example, student work samples could be used to show understanding well beyond what is normally expected from students at their grade level as part of making a case that the interventions were effective.

[4] Researchers typically use the term *validity* when describing the appropriateness of an instrument for a study, and *reliability* to describe the likelihood that repeated use of an instrument (or the application of that instrument by different researchers) would yield consistent results.

[5] This situation is similar to selection bias, but in this case, it is the patterns of response rather than the initial designation of groups that creates the problem, which researchers quite logically refer to as *non-response bias*.

treatment group dropped out, a study might falsely conclude that a treatment was effective. On the other hand, if nearly all of the treatment group teachers responded, but only the best prepared teachers in the comparison group responded, a study might mistakenly conclude that an effective treatment did not work. In deciding how much confidence to put in the findings, look for evidence that the respondents were in fact similar to the non-respondents. And, of course, you should be concerned if no information about response rates is provided.

Look for clues about the appropriateness of the data analysis.

Although determining whether researchers used appropriate methods for analyzing their data requires specialized knowledge, consumers of research who are not trained in research methodology can look for clues about the trustworthiness of the analyses. For example, researchers who analyzed qualitative data such as video recordings or interview transcripts should describe how they decided what to look for, how data analysts were trained to know it when they saw it, and their strategies for ensuring they did not overlook evidence that was different from what they expected.

Results of quantitative analyses are more likely to be sound if the researchers conducted a small number of carefully selected statistical tests; running many tests using the same data would increase the likelihood that differences that are actually the result of random variations in data will be mistakenly labeled as statistically significant. Researchers often talk about a result being significant at the "0.05 level," which means there is less than a 5 percent probability that the result was actually due to chance. If a study compared treatment and control group results on an assessment by considering differences in results by gender, race, school size, parent background, teacher credentials, and teacher experience, without some method of adjusting the analysis to account for the multiple comparisons, the probability of mistaking *something* that is actually due to chance as statistically significant rises to more than 25 percent. Results of quantitative analyses are particularly questionable if only a subset of the results are provided, suggesting that there may be some "cherry picking" going on.

Sample size also comes into play. If samples are small, interventions that really do work might appear to be ineffective because the studies did not have sufficient statistical power to detect their impact. On the other hand, if samples are very large, then just about any difference will be statistically significant, whether or not it is educationally meaningful. Consumers of research should also be alert to a related problem—sometimes studies assign individual classes (or teachers or schools) to treatment or comparison groups, but then base their analyses on the number of students in these groups, resulting in an inflated sample size and possibly leading to incorrect conclusions.

Because statistical significance depends on sample size and does not convey whether a result is practically important, researchers are increasingly being asked to include "effect sizes" in their reports. Effect sizes are intended to provide, in meaningful terms, a measure of how large a change in outcome is associated with an intervention, or how strongly two variables are related.[6] For instance, whether statistically significant or not, in most cases, it will not make sense to spend resources if the expected difference for doing so is very small, e.g., increasing the average score on an assessment from 70 percent to 70.5 percent.

---

[6] Effect sizes of about 0.20 are typically considered small, 0.50 medium, and 0.80 large. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Use similar criteria in assessing the trustworthiness of research syntheses.

Research syntheses can be a boon to consumers of research, saving the time it would take to examine a large number of individual studies, but only if the syntheses are designed, implemented, and communicated well. Like reports of single studies, research syntheses provide information about data collection, in this case including how they identified potentially relevant studies and the criteria they used to select the studies that were reviewed. Readers should look for evidence that a research synthesis took both effect sizes and the quality of the individual studies into account, e.g., applying appropriate standards of evidence and giving more credence to the findings of well-designed studies rather than simply counting the number of studies that reported a particular finding.

Section 2: What are the implications for my context?

Credible results from a single study might lead you to reflect on the implications for your context, but typically you would wait to "do" something until you had more evidence in support of a particular finding, especially if the action involved would be time-consuming or expensive. If multiple studies with different research designs, and different strengths and weaknesses, reach similar conclusions, you can have more confidence in the findings, but you still need to consider whether the findings are likely to apply to your context. Earlier, as part of judging whether a study's findings were trustworthy, we suggested some characteristics that might matter in judging whether treatment and comparison groups were initially equivalent, such as teacher background and experience, school contexts, demographics, and prior performance results. The same kinds of considerations are important in assessing the extent to which particular findings might be applicable in your situation. For example, a mathematics supervisor might need to consider whether a professional development approach that has been shown in multiple studies to be effective with middle and high school mathematics teachers would likely also be effective with elementary school teachers. And you would want to think long and hard about pursuing an intervention that would not be well received in your context, perhaps because it sounds too much like something that was problematic or ineffective in the past.

If the findings seem applicable, the next question is whether you will be able to in fact apply them. First, you need to consider whether you have enough information about the interventions that were found to be effective in order to replicate them. Rarely are interventions described in detail in research articles; the emphasis tends to be on study design, data collection, analysis, and results. Consequently, you may know that something was effective, but not really know what "it" was. However, if the intervention is based on a published resource, such as professional development materials, and it was shown to be effective in multiple implementations by different providers, you can have reasonable confidence that the intervention is a robust one, even in the absence of details about how it was implemented in any particular instance.

Finally, in considering the implications of research results, you need to take into account the capacity and resources available to you. A key question is whether you will have the time available to implement a new intervention. Many teacher professional development programs require substantial time for teacher participation and even more for those who will lead the program. Another consideration is how many leaders, with what expertise, are needed to carry out an intervention. An intensive one-on-one peer coaching program that had great success in multiple implementations in multiple contexts might be very appealing to you. However, you will not get very far with that program unless you have access to well-prepared individuals who can serve as coaches, or can

develop them. In this situation, you would do far better to look to see if there is a scaled down version of the approach, or another approach with solid evidence of effectiveness that is feasible for implementation in your context.